Principles of data organization

Database

database	a collection of related structured information about entities
file	a collection of records
record	a set of fields
field	a single characteristic of an entity
character	a symbol used in data field

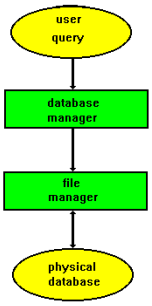
Example of a Genbank entry

```
LOCUS       VIBHALUXA    3141 bp    DNA             BCT       15-FEB-1996
DEFINITION  V.harveyi luciferase alpha and beta subunit (luxA and luxB) genes,
            complete cds.
ACCESSION   M10961 M13494
NID         g155174
KEYWORDS    luciferase.
SOURCE      Vibrio harveyi DNA.
  ORGANISM  Vibrio harveyi
            Eubacteria; Proteobacteria; gamma subdivision; Vibrionaceae;
            Vibrio.
REFERENCE   1  (bases 1 to 1838)
  AUTHORS   Cohn,D.H., Mileham,A.J., Simon,M.I., Nealson,K.H., Rausch,S.K.,
            Bonam,D. and Baldwin,T.O.
  TITLE     Nucleotide sequence of the luxA gene of Vibrio harveyi and the
            complete amino acid sequence of the alpha subunit of bacterial
            luciferase
  JOURNAL   J. Biol. Chem. 260 (10), 6139-6146 (1985)
  MEDLINE   85207595
REFERENCE   2  (bases 1745 to 3141)
  AUTHORS   Johnston,T.C., Thompson,R.B. and Baldwin,T.O.
  TITLE     Nucleotide sequence of the luxB gene of Vibrio harveyi and the
            complete amino acid sequence of the beta subunit of bacterial
            luciferase
  JOURNAL   J. Biol. Chem. 261 (11), 4805-4811 (1986)
  MEDLINE   86168191
```

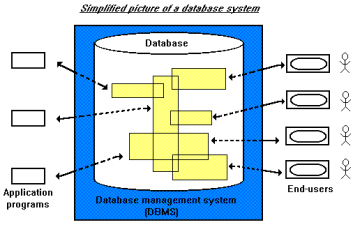
Example of a Genbank entry

```
FEATURES             Location/Qualifiers
     gene            707..1774
                     /gene="luxA"
     CDS             707..1774
                     /gene="luxA"
                     /codon_start=1
                     /product="luciferase alpha subunit"
                     /db_xref="PID:g155175"
                     /transl_table=11
                     /translation="MKFGNFLLTYQPPELSQTEVMKRLVNLGKASEGCGFDTVWLLEH
                     HFTEFGLLGNPYVAAAHLLGATETLNVGTAAIVLPTAHPVRQAEDVNLLDQMSKGRFR
                     FGICRGLYDKDFRVFGTDMDNSRALMDCWYDLMKEGFNEGYIAADNEHIKFPKIQLNP
                     SAYTQGGAPVYVVAESASTTEWAAERGLPMILSWIINTHEKKAQLDLYNEVATEHGYD
                     VTKIDHCLSYITSVDHDSNRAKDICRNFLGHWYDSYVNATKIFDDSDQTKGYDFNKGQ
                     WRDFVLKGHKDTNRRIDYSYEINPVGTPEECIAIIQQDIDATGIDNICCGFEANGSEE
                     EIIASMKLFQSDVMPYLKEKQ"
BASE COUNT      883 a    665 c    741 g    852 t
ORIGIN      1 bp upstream of EcoRI site.
        1 gaattcacca tgacgacggg caaaaatagt ttgtgcactg tttatcactg gctgcagacc
       61 aagggcacac aaaacattgg cttgattgcg gcaagtctct cagctcgtgt cgcctatgaa
      121 gttatctctg atctggagct gtcttttctg attactgcgg ttggtgtggt gaacttgcgt
      181 gacacactag aaaaagcgct tggttttgat tacctcagtt tgcctatcga tgagctacca
          ....
```

Database Organization



Database Management System (DBMS)



Four major components of DBMS:

Data * Hardware * Software * Users

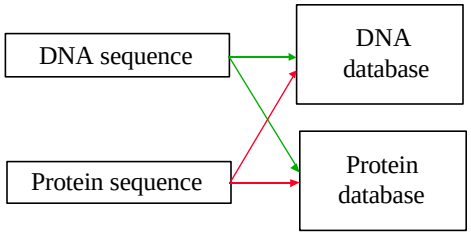
Data Model

- A named logical unit (record type, data item)
- Relationships among logical units

Relationships among logical units

- one to one
- one to many
- many to one

DNA vs. Protein searches



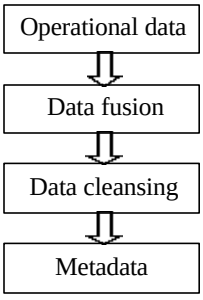
DNA sequence - DNA database

- larger databases
- more random hits
- simpler scoring functions
- missing hits (similar proteins encoded by different DNAs

Database administration

- Redundancy eliminated
- Inconsistency avoided
- Data shared
- Standards enforced
- Security applied
- Integrity maintained
- Requirements balanced

Data Warehouse



Data Mining

- Data mining is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules
- Common data mining tasks
  - Classification
  - Estimation
  - Prediction
  - Affinity Grouping
  - Clustering
  - Description

Knowledge Discovery

- Directed and Undirected KD
- Directed KD
  - Purpose: Explain value of some field in terms of all the others
  - Method: We select the target field based on some hypothesis about the data. We ask the algorithm to tell us how to predict or classify it
  - Similar to hypothesis testing (e.g., in regression modeling) in statistics

Knowledge Discovery

- Undirected KD
  - Purpose: Find patterns in the data that may be interesting
  - Method: clustering, affinity grouping
  - Closest to ideas of machine learning in artificial intelligence
- Comparison
  - UKD helps us to recognize relationships & DKD helps us to explain them

Classification

- Classifying observations into different categories given characteristics

Prediction

- Rules that explain how to predict a future value or classification, given characteristics

Estimation

- Rules that explain how to estimate a value given characteristics

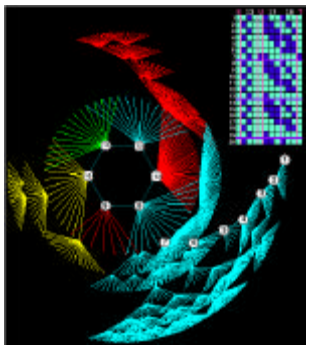
Affinity Grouping

- Grouping by relations (not by characteristics)

Clustering

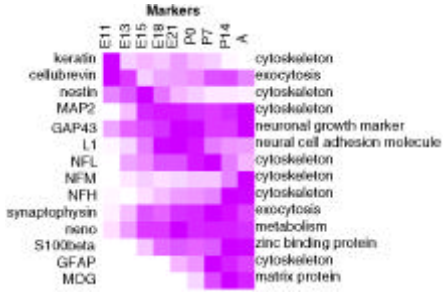
- Segmenting a diverse population into more similar groups
- In clustering, there are no pre-defined classes and no examples. Records are grouped together by some similarity measure.

Pattern visualization



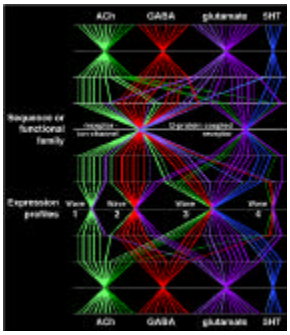
Basin of attraction of 12-gene network model. (Somogyi & Sniegoski, *Complexity* **1**:45,1996)

Gene expression data analysis



Temporal expression patterns for genes expressed in rat spinal cord
(Wen et al. *Proc Natl Acad Sci USA*, **95**:334, 1998.)

Expression clusters and gene families



Neurotransmitter receptors follow particular expression waveforms according to ligand and functional class.
Agnew, *Science*, **280**:1516,1998)